Direct nanopore sequencing of mRNA reveals landscape of transcript isoforms in apicomplexan parasites

V Vern Lee[1,2], Louise M. Judd[3], Aaron R. Jex[2,4] Kathryn E. Holt[3,5], Christopher J. Tonkin[2,6], Stuart A. Ralph[1].

1. Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Melbourne, Victoria, 3010, Australia

2. The Walter and Eliza Hall Institute of Medical Research, Parkville, Melbourne, Victoria 3052, Australia.

3. Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria, Australia

4. Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, Victoria, Australia

5. London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK

6. Department of Medical Biology, The University of Melbourne, Melbourne, Victoria 3052, Australia.

Abstract:
Alternative splicing is a widespread phenomenon in metazoans by which single genes are able to produce multiple isoforms of the gene product. However, this has been poorly characterised in apicomplexans, a major phylum of some of the most important global parasites. Efforts have been hampered by atypical transcriptomic features, such as the high AT content of Plasmodium RNA, but also the limitations of short read sequencing in deciphering complex splicing events. In this study, we utilised the long read direct RNA sequencing platform developed by Oxford Nanopore Technologies (ONT) to survey the alternative splicing landscape of *Toxoplasma gondii* and *Plasmodium falciparum*. We find that while native RNA sequencing has a reduced throughput, it allows us to obtain full-length or near full-length transcripts with comparable quantification to Illumina sequencing. By comparing this data with available gene models, we find widespread alternative splicing, particular intron retention, in these parasites. Most of these transcripts contain premature stop codons, suggesting that in these parasites, alternative splicing represents a pathway to transcriptomic diversity, rather than expanding proteomic diversity. Moreover, alternative splicing rates are comparable between parasites, suggesting a shared splicing machinery, despite notable transcriptomic differences between the parasites. This work highlights a strategy in using long read sequencing to understand splicing events at the whole transcript level, and has implications in future interpretation of RNA-seq studies.

Introduction

Transcriptomic analyses have been central to insights into the biology and pathogenesis of eukaryotic pathogens. The best-characterised eukaryotic pathogen transcriptomes are those of the phylum Apicomplexa. This phylum includes some of the most important parasites impacting human and veterinary health, such as *Plasmodium* and *Toxoplasma*. *Plasmodium* is the causative agent of malaria, a devastating parasitic disease infecting over 200 million individuals and killing 400,000 each year [1]. *Toxoplasma* causes toxoplasmosis, a widespread

zoonoses that primarily impacts immunocompromised, young, and pregnant individuals [2], and is thought to infect a third of the world's population [3]. The pathogenesis of apicomplexan infections is intimately linked to the parasites' life cycles. The life cycle of most parasitic apicomplexans is complex, involving multiple differentiated forms and hosts, and this requires reprograming of the parasite transcriptome.

Early transcriptomic experiments sought to utilise techniques such as microarrays and Sanger sequencing of cDNA or EST libraries to understand changes in gene expression that define the pathogenesis of the parasites. These studies reveal that the timing of appearance and abundance of individual mRNAs follow developmentally distinct patterns [4], even for many predicted housekeeping genes. For example, the expression of the actin gene family in *P. falciparum* is developmentally tuned, with actin I primarily transcribed in asexual intraerythrocytic life stages while actin II is primarily present in sexual stage parasites [5, 6]. Unusually, however, there is a poor correlation between protein and mRNA expression profiles for many genes in parasitic apicomplexans [7]. In one experiment, Foth and colleagues found widespread discrepancies between temporal expression patterns of proteins and transcripts in *P. falciparum* [8]. Such discrepancies suggest that substantial post-transcriptional regulation occurs within these parasites. Indeed, with the advent of RNA-seq, more recent studies now show that multiple layers of gene expression are required for parasite life progression, through transcriptional, post-transcriptional, and epigenetic control mechanisms [9-11].

RNA splicing provides one such source of co- and post-transcriptional regulation. In this process, introns are removed from the pre-mRNA and the exons retained to form one contiguous molecule that is then translated by the ribosome. However, for complex mRNAs, alternative splicing either of untranslated regions, or the exonic chain, can add additional complexity. Through this process, pre-mRNA species can be differentially spliced, to create multiple distinct mature mRNAs from a single gene. This can alter regulation of the gene, for example by removing small-RNA binding sites [12] or diversify the proteome, as individual genes may encode multiple protein isoforms with altered structure or function [13].Indeed, proteomic analyses have revealed widespread protein isoforms arising from single genes, corresponding with varying activity, stability, localisation and post-translational modifications [14, 15]. With advances in genome and transcript sequencing, it has become apparent over the last decade that alternative splicing of pre-mRNA occurs to a great extent. For example, more than 95% of human genes are alternatively spliced, and many transcript isoforms are specific to tissues or cellular states [16]. Such observations suggest that RNA diversity is more complex than previously appreciated [17].

Although alternative splicing appears to play a major (though debated) role in post-transcriptional control in metazoans, the process in less understood in apicomplexans. Studies have identified apicomplexan genes with crucial alternative splicing outcomes [18]. For example, alternative splicing is required for attaching a protein trafficking pre-sequence onto two adjacent gene coding sequences [19], and normal multi-organellar targeting of the *P. falciparum* cysteinyl tRNA synthetase, which is essential for parasite survival [20]. Nonetheless, there is little other study of alternative splicing in this phylum. Understanding diversity of parasites transcripts is crucial for drug and vaccine development because certain putative target genes may produce isoforms that escape the intervention. This has been postulated for the chloroquine resistance transporter gene of *P. falciparum* (*Pf*CRT) in clinical isolates, though the role of the splice variants remains unclear [21]. In other organisms, there is some evidence showing that essential genes are more likely to have alternatively spliced transcripts compared to non-essential genes [22, 23]. This has not been explored in

apicomplexans but highlights further considerations for investigating drug targets and interventions.

The lack of data for apicomplexan gene isoforms is a major obstacle to dissecting the complexity of transcript outcomes. Traditionally, transcriptomic studies employing RNA-seq have relied on short read technologies such as Illumina, 454 and Ion-torrent [24]. Despite the power of very high sequencing depth and low error rates, the short reads present a limitation in that simultaneously occurring alternative-splicing events within individual transcripts cannot be unambiguously detected or linked. Previously-developed computational methods for full-length transcript assembly from short read sequencing data are often computationally intensive, and can produce ambiguous or conflicting results between different algorithms [25]. In addition, sequencing on cDNA strands amplified by PCR has a propensity to introduce biases in relative transcript abundances and rare isoform identification [26]. Hence, it is difficult to draw functional relationships between simultaneous alternative splicing events and observable phenotypes. In apicomplexan parasites, simultaneously occurring alternative splicing events within a specific transcript isoform do occur [27]. However, the studies that unearthed these transcript isoforms relied on cDNA probes and reverse transcription PCR, and the wider extent of this phenomenon is unknown.

Recently-developed third generation sequencing platforms, such as those developed by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), are capable of producing significantly longer reads at the single-molecule level. These technologies have been used in various applications such as resolving genomic and transcriptional landscapes [28, 29], single cell transcriptome sequencing [30], and DNA or RNA methylation pattern profiling [31, 32]. PacBio has recently been used to generate an amplification-free transcriptome from *Plasmodium falciparum* cDNA, which has helped to elucidate transcriptional start sites and to improve annotation of the 5'and 3' UTRs [33]. Unlike most other sequencing platforms, a notable characteristic of ONT sequencing is the ability to directly sequence native RNA [34]. With this methodology, each read represents a complete molecular transcript, which could thus significantly resolve weaknesses of amplification-based RNA-seq. In particular, each spliced isoform need only be counted as individual reads, as opposed to complex assignment and assembly of multiple spliced reads. Furthermore, due to differences between DNA and RNA molecules, contaminating DNA sequences cannot be correctly base-called after sequencing and so are easily discarded [35]. Recently, several studies have successfully applied single molecule, long-read sequencing to identify a high number of novel transcript isoforms [28, 36, 37]. However, these studies have also identified several caveats including a reduced throughput and high error rates.

In this study, we evaluate the ability ONT direct RNA sequencing to characterise the alternative splicing landscape of two parasitic apicomplexans, *T. gondii* and *P. falciparum.* Our analyses show that alternative splicing, particularly intron retention, is extensive throughout the transcriptome, with most multi-exon genes having some degree of intron retention, and some genes only rarely producing transcripts with all introns removed. The long reads produced from ONT sequencing showed that most of these alternative splicing events are likely non-productive in protein-coding capacity, but may provide an additional layer of gene expression regulation.

Results

Direct RNA sequencing of *T. gondii* and *P. falciparum* allows the detection of full-length transcripts

We generated ONT sequencing reads of polyA-selected RNA from asynchronous *T. gondii* (Pru) tachyzoites and *P. falciparum* (3D7) mixed asexual intraerythrocytic stage parasites. Mixed-stage cultures were used to maximise transcript diversity for these samples. For *T. gondii*, we obtained a total of 310,813 reads corresponding to about 500 million bases (Mb). For *P. falciparum*, we obtained a total of 456,098 reads, corresponding to about 300 Mb of data. Although the *P. falciparum* sample yielded fewer sequenced bases, we estimate the theoretical gene coverage for both parasite samples to be similar at 25-26 fold due to differences in gene number and length. Using minimap2 [38], we successfully mapped 78.90% of the *T. gondii* reads, and 44.48% of the *P. falciparum* reads to the parasite genomes. We analysed the quality of the sequencing reads using FASTQC, and found consistently high-quality scores over the length of reads, with no drop-off in quality even at reads of 10 Kb (Fig 1B). This is important because base quality scores generally correlate with read accuracy to the reference sequence. That is indeed the case for our dataset (supplementary figure S1). We used AlignQC to estimate the base-call error rate of the transcript reads based on aligned segments and found, on average, an error rate of 19-20% for both parasites. The read-length distributions of the mapped and unmapped data (Fig 1A) show that the mapped reads are predominantly longer than the unmapped reads, with some read lengths exceeding 10 kb. As expected, there was a sharp increase in unmapped read counts at the ~1.35 kb length (Fig 1A) corresponding to yeast enolase 2, a calibration standard added during the library preparation. We calculated the average mapped read lengths to be ~1.9 k and 1.3k bases for *T. gondii* and *P. falciparum* respectively, well within previous range estimates of predicted transcript lengths of both organisms [39].

To evaluate the capability of ONT sequencing to generate full-length transcript reads, we re-mapped the reads to the parasites' annotated transcriptome and calculated the fraction of full-length transcripts per gene. We define full-length transcripts as reads that cover more than 95% of the predicted canonical transcript based on the annotation file. As illustrated in the Figure 1C, many transcripts were observed to have full-length reads, more so for the *P. falciparum* data. In *T. gondii*, 1117 genes have 75% or more of their corresponding reads that were considered full length. In *P. falciparum,* this number is 1835 genes. The difference can be attributed to the absence of 5' and 3' UTR annotations for *P. falciparum*, which results in the underestimation of predicted transcript lengths. In both cases, the fraction of mapped reads corresponding to full-length transcripts fell with increasing transcript length, independent of expression levels, which is consistent with read truncation disproportionately affecting the longest reads. To better understand the overall distribution of transcript coverage, we calculated the average coverage of the transcripts per gene. Again, there is a general decrease in transcript coverage with increasing transcript length (Figure 1C). However, many of the genes retain a high level of transcript coverage, even when there is a low fraction of full length reads. For example, in *T. gondii*, genes with predicted transcript lengths of 3 kb or longer only had an estimated 12% of their reads that were considered full length on average, even though the average read coverage for those genes is 50.64%. The overall average transcript coverage is calculated to be over 60% in both parasites (*T. gondii*: 64.73%, *P. falciparum*: 80.51%). Together, the data indicate a generally high proportion of full- or near full-length transcript reads.

ONT sequencing is comparable with traditional RNA-seq for quantifying gene expression levels

To investigate the utility of ONT data to measure transcript abundance, we computed read-count correlations between our ONT dataset and reanalysed published Illumina-based RNA-seq datasets on comparable parasite samples. In theory, ONT RNA sequencing reads directly correspond to complete transcripts and so quantifying the expression of genes can be done by simple counting of the assigned reads. This is dissimilar to traditional short read RNA-Seq which necessitates a further normalisation step (e.g., reads or fragments *per kilobase of transcript* or *transcripts per million*) to account for the higher number of reads that would be generated from longer transcripts. For *T. gondii*, we used Illumina datasets from a closely related strain (ME49) due to a relative lack of comprehensive datasets for the Pru strain. Strikingly, we observe strong positive correlations between the ONT and Illumina datasets (Fig 2A & B), regardless of mapping to the transcriptome or genome (Spearman's rho = 0.81 for transcriptome, 0.87 for genome). For *P. falciparum* we correlated the mixed stage ONT dataset with individual datasets from three main developmental stages (rings, trophozoites and schizonts), and a final combined dataset. In all cases, we found moderately positive correlations between the datasets. As expected, the correlation is higher in the later stages (supplementary figure S2; Spearman's rho = 0.47 rings, 0.57 for trophozoites, 0.63 for schizonts), and the highest with the combined dataset (Figure 2A/B; Spearman's rho = 0.64 for transcriptome, 0.68 for genome), a reflection of mRNA abundance in these different stages.

For both parasites, a higher number of gene transcripts were detected in the Illumina datasets than in the ONT data (supplementary table TS1). This is expected, given the greater sequencing depth that was obtained from the Illumina runs. The theoretical average fold coverage from our Illumina data is estimated to be 750 and 1000 times of the *T. gondii* and *P. falciparum* genes respectively (compared to 25-26 fold for the ONT data). Notably however, abundant non-mRNA species, particular ribosomal RNA, were detected in the Illumina datasets. This is absent in the ONT datasets, possibly because of differences in polyA RNA purification methodologies or biases due to PCR amplification. We further evaluated the ONT transcriptomes for genome coverage completeness as shown in Figure 2B. The read coverages from the ONT and Illumina reveal no significant bias towards a particular region of the nuclear genome.

Intron retained transcripts are prevalent and generally non-productive

A major goal of mature full-length transcript sequencing is the identification of splicing isoforms. Alternative splicing can be broadly classed into four types: intron retention, alternative 3′ splice site selection, alternative 5’ splice site selection, and exon skipping [40]. Of these, intron retention is the least studied form of alternative splicing despite the numerous studies implicating the significance of the event [41-43]. This is in part due to the limitations of short read sequencing, but also the relatively long and low-complexity introns in metazoan genomes, which impose limitations on sequencing and assembly. For example, the intronic sequence in the human genome is several magnitudes longer than the length of the exonic sequence [44, 45]. In contrast, the compact genomes of *Plasmodium* and *Toxoplasma* both have gene models that have similar or longer exon lengths than introns [46, 47], so reads that span multiple entire introns are quite achievable for these organisms.

To monitor levels of intron retention, we identified junctions and reads that overlapped annotated intronic regions of a gene based on the annotated coordinates using FeatureCounts [48], and tallied the proportion of reads mapping to that intron to the total reads for the same gene. Proportion scores are represented using the metric *percent intron retention* (PIR). Based

on this analysis, 17.65% of the mapped reads were considered to have intron overlaps for *T. gondii*, and 4.54% for *P. falciparum.* We further filtered out junctions without a minimum overlap of six bases to exclude artefacts generated by read errors, and excluded genes that were not supported by a minimum coverage of three reads. The distribution of PIR scores per gene (Figure 3A) reveals an overall skew towards low proportions of intron-overlapping reads. This is as expected given the propensity for a dominant canonical transcript [49]. However, we identify a strikingly high number of genes that retain a high level of intronic regions. Using a threshold of 10% PIR, we identified a total of 3229 genes for *T. gondii* and 978 genes for *P. falciparum* that have intronic reads within their transcripts (supplementary table TS2). Moreover, for around 29.82% (963/3229) of the *Toxoplasma* genes, and 19.63% (192/978) of *Plasmodium* genes, 50% or more of the reads retain at least one intron. Unusually, there are a considerable number of genes where none of the transcripts appears to have all of their annotated introns removed. We manually investigated these cases further and found major differences between the transcripts and gene model in most cases, suggesting that these highlight genes with incorrectly annotated structure. A couple of examples are outlined in supplementary figure S3. Most of these genes are annotated as hypothetical proteins, highlighting the potential of ONT sequencing to validate gene models.

The most extreme cases of conflict between the junctions we detect and canonical gene models often highlight potential annotation errors, but there are still a strikingly high number of genes where genuine introns are retained in a high (>50%) proportion of transcripts (*T. gondii*: 808 genes, *P. falciparum*: 162 genes). Additionally, the differences between the two parasites are striking. In many organisms, those transcripts with the most introns are those that are more likely to retain at least one or more introns [50]. This is partially supported in our analysis, where we observe a higher level of overall intron retention for *T. gondii* (which has 4.5 introns per gene on average) than *P. falciparum* (which has only 1.5 introns per gene on average). A possible explanation for this relationship is thus that both organisms have a similar level of intron retention for any given junction, and the higher average intron number in *T. gondii* genes results in more overall intron retention per gene. To examine this, we calculated PIR scores at the individual junction level, rather than per-gene level, and normalised the count of each PIR value to the proportion of total junctions within each organism. The analysis reveals that after correction for intron number there is virtually no difference in the distribution of intron retention levels between parasites (Fig. 3B). In other words, individual *T. gondii* junctions are no more likely to experience intron retention than *P. falciparum* junctions. We further tested whether intron number was the major predictor of intron retention at the gene level by looking at the correlation between the number of introns per gene and levels of intron retention. Interestingly, we only obtained poor or moderate positive correlations in all datasets (supplementary figure S4; Spearman's rho = 0.28 for *T. gondii*, 0.48 for *P. falciparum*, 0.39 for pooled). This correlation does not significantly improve even when we restrict our analysis to higher (≥ 10 reads) expressed genes. This suggests that while intron number is associated with increased level of intron retention, it is not the main determinant of whether a gene retained at least one intron.

By taking advantage of the full-length reads made possible by ONT, we are able to predict the protein-coding productivity of the alternate transcripts. We performed productivity analysis on full-length intron-retained reads using the FLAIR [51] pipeline, which corrects and defines unproductive transcripts as transcripts with a termination codon that is 55 nucleotides or more upstream of the 3'-most splice junction. The rationale for this definition is based on previous evidence suggesting that only premature terminating transcripts following that 55-nucleotide rule mediate an effect on mRNA turnover [52]. This is a conservative estimate of productivity

as it does not consider intron retention within the 3'-most splice junction. The Flair method identified over 70% of the intron-retained reads to be non-productive for either parasite (Fig 3C), suggesting that the high level of observed intron retention only rarely corresponds to alternative protein products, and that most intron-retaining transcripts may instead be targets for nonsense mediated decay. Intron retention is known to fine-tune protein expression through this pathway in mammalian systems [53]. A related prediction from other studies [54] is that the most highly expressed transcripts should have low levels of intron retention. In our analysis, we do observe a negative relationship between intron retention and gene expression levels (Fig. 3D). There is a relatively high variance for this, more so for genes with lower expression levels. This is likely due to the limitation in precision for the lower sequencing depths. For example, intron retention occurring in 10% of transcripts for a given gene will not be precisely measured for a gene for which only five reads are available. We circumvented this by classifying the transcripts into bins of equal read number based on expression and quantifying global intron retention levels within each bin. Again, we observe a negative relationship between intron retention and expression levels (supplementary figure S5). To investigate the functional significance of this, we further analysed the genes for Gene Ontology (GO) enrichment. Here, we only considered genes with a minimum coverage of 10 reads to increase precision. The analyses reveal the consistent enrichment of genes with functions associated with the ribosome when there are lower levels of intron retention across both parasites (supplementary table TS3). This association has been previously observed in other organisms [55], though its basis is unknown. We also tested whether intron retention correlated with essentiality based on previous functional genomic screens [56, 57], and found no significant relationships (supplementary figure S6).

To validate the identification of intron retention events, we looked at whether retained introns apparent in the ONT data were directly supported by Illumina RNAseq data. We normalised read counts by junction length and only considered intronic data that spanned the full junction. Based on the analysis, 77.88% for *T. gondii* and 87.37% for *P. falciparum* of the intronic junction reads flagged from the ONT datasets were supported by Illumina reads. However, we also noted that some alternative splicing events, particularly the lower frequency ones, failed to be captured by ONT sequencing compared to the Illumina dataset (Fig 3E). This is again likely due to the limitation in read depth in the ONT dataset. Based on our sequencing of polyA tailed material combined with previous kinetic studies [58, 59], we do not expect the intron retained transcripts to simply be unprocessed transcripts. To confirm this, we looked for evidence that each transcript had at least been partially processed. On average, 92.76% of multi-intron genes identified as having intron retention within their transcripts had at least 1 junction which was canonically spliced in all the transcripts, demonstrating that cannot be attributed to sequencing of pre-mRNA.

Alternate junction splicing is often proximal and non-productive

Having previously identified intron-retained, read junctions using an annotated gene model approach, we used RSeQC to identify and quantify the other three classes of alternative splicing read junctions (exon skipping, 5'-, and 3'-splice site change) based on a similar methodology. Levels of alternative spliced junctions are calculated as the proportion of alternate junction over the total junction reads, and are represented using the metric percent spliced (PS). Here, we filtered out junctions unsupported by a minimum coverage of three reads. Using the same threshold as before (≥ 10%), we identified a total of 1138 genes for *T. gondii,* and 168 genes for *P. falciparum,* where one or more of their junctions exhibited alternative 5'/3' splice site selection or exon skipping (supplementary table TS4_1 & 2). Remarkably, these aggregate numbers are lower than those we calculated for intron retention alone. Combining the datasets,

intron retention accounts for 60-68% of alternatively spliced genes identified, alternate 5' junction and 3' junction splicing for 13-19% and 6-11% respectively, and exon skipping for less than 3% (Fig. 4). The rest of the junctions flagged in the analysis defy easy categorisation due to major mismatches between the RNAseq data and the annotated gene model. Subsets of genes were also found to have multiple alternative splicing type events within their transcripts as observed in supplementary figure S7, though there does not appear to be a particular functional trend.

Having identified junctions subject to alternative splicing, we then quantified what proportion of the transcripts produced at those junctions represented the non-canonical isoform. For alternative splicing involving 5' or 3' changes and for intron retention, substantial proportions of isoforms were represented by the alternative transcript, but the median abundance remained below 50% (data shown in Fig. 5A). However, whilst exon skipping was a relatively rare event across the genome (Fig. 4), for those genes where exon skipping did occur, it represented a higher proportion (approximately two-thirds) of transcript isoforms for those genes than observed for the other forms of alternative splicing (Fig. 5A). We had previously published a list of genes with alternative splicing excluding intron retention in *T. gondii* (RH) using our in-house built computational tool– JunctionJuror [60]. A limitation of the tool over our current methodology is that the information about junction types is ultimately lost. Despite the differences in methodologies and parasite strain, we compared our list of genes with alternative splicing excluding intron with the previous list and found a moderately high degree of overlap (57.46 %) (supplementary table TS5). This intersect is double that is expected by chance alone and may represent genes with a steady state alternative splicing of higher confidence.

To further explore consequences of alternate 5' or 3' junction splicing events, we investigated the length distribution of the change in intron length. We found a surprisingly high proportion (~50%) of alternate 5'/3' splicing to occur proximal (<6 bases) to the expected canonical site. We graphed the distribution of splicing change positions in Figure 5B and found a substantial spike of splicing changes occurring at the position four bases inside the canonical intron boundary. We tested these junction reads for productivity for a subset of 50 of these "near-miss" alternate splice events, and found all reads to prematurely terminate (unsurprisingly given the necessary frame-shift). This striking over-representation of isoforms departing from the canonical model by specifically four bases has been previously identified in metazoans [61]. Very small movements in splice site usage have been described as junction wobbling, and this has been proposed as minor splicing noise, or alternatively, as an additional mechanism of regulation through the NMD pathway [61], although the reason for the specific peak of AS four bases away from the canonical junction is unknown.

Discussion

Despite the apparent significance of alternative splicing in metazoans, very little is known about the process in apicomplexans. A number of targeted experiments highlight single splicing events and their impact on parasite survival, but global splicing networks have been poorly described. Untargeted RNA-seq experiments have mainly focused on whole transcriptome assembly and/or gene expression. Those studies that do monitor alternative splicing reveal its occurrence in multiple genes and stages, but the extent of these events and their phenotypic significance remains unknown. The lack of a robust methodology in defining transcript isoforms from short read data is a particular challenge in dissecting whole-transcriptome splicing. In this study, we investigated whether ONT direct RNA sequencing

could be used to explore the splicing landscape of two apicomplexan parasites- *T. gondii* and *P. falciparum.*

To our knowledge, ONT direct RNA sequencing has not been previously described in apicomplexans and so our first objective was to evaluate the capability of ONT sequencing in generating sequencing reads from these parasites. We successfully obtained high quality sequencing reads for both parasites that were comparable to that previously described in the literature for other organisms [62, 63]. In particular, we obtained read lengths that exceeded 1kb on average, many of them predicted to represent full-length or near full-length transcripts. Interestingly, although we obtained a higher number of sequencing reads for *P. falciparum,* the mapping of the reads was suboptimal compared to that of *T. gondii*. Repetitive DNA sequence motifs are characteristic of many large eukaryotic genomes and this has been known to complicate the mapping of reads that cannot be confidently assigned to these particular regions [64]. In theory, long read sequencing mitigates this problem because long enough reads should unambiguously match a unique site on the genome, irrespective of low complexity or repeat sequences. However, the genome of *P. falciparum* is particularly AT-rich (~82%) with numerous regions of extreme low complexity [65]. Thus, we may expect some reads, particularly the shorter ones, to fail the mapping parameters. Indeed, as indicated above, reads that fail to map to the genome tended to be shorter than reads that do. This is further exacerbated by the high error rate produced from ONT sequencing. Based on previous experiments, the per base error rate of direct RNA sequencing using ONT is 10-20% [51, 62]. In our dataset, we estimated the error rate to be around 20%. This may have further contributed to the poorer mapping, though we do not expect the high error rate to significantly impact our study because the main analysis is focused on splice connectivity, rather than base sequences. As has happened for ONT DNA sequencing, we are likely to see significant improvements in read and mapping accuracy of RNA sequences as improvements are made to the flow cell and base caller. A study carried out by Runtuwe and colleagues is an elegant example, where ONT DNA sequencing on targeted *P. falciparum* genes yielded a mapping percentage that improved from 57.86% to 92.46% with improved chemistry of the flow cell, and upgrades to the base-calling algorithms [66].

The quantification of gene expression is one important goal of RNA-seq. Traditional, short-read sequencing requires the generation and amplification of complementary DNA (cDNA) which can introduce artefacts and biases. Transcriptional amplification or repression is a commonly-overlooked bias [67], where the levels of global mRNA, rather than specific mRNA, may be variable between different samples. Thus, using a standard amount of total RNA, as is commonly done, can mask actual detection of specific mRNA levels, even after normalisation [67]. Direct RNA sequencing allows these caveats to be bypassed because a standard amount of isolated mRNA instead is used as the sequencing material. However, because there is no amplification step, direct RNA sequencing is limited by the amount of mRNA that can be practically obtained and used in the sequencing process. Without sufficient sequencing material, it can be difficult to achieve the high levels of sequencing depth that is needed to analyse gene expression [68]. In line with the literature, our analysis shows that the current protocol for ONT direct RNA sequencing is comparable to Illumina for quantifying gene expression in the organisms we analysed. It can be noted however that sequencing depth remains the main limitation of ONT sequencing in our study. The reduced throughput and sequencing depth from ONT sequencing compared to Illumina sequencing means that genes or transcript isoforms with low expression may not be captured.

Several previous analyses have reported differences in alternative splicing types and levels among different organisms [69-71]. Notably, the increase in intron number and its retention correlates strongly with multicellular complexity (as defined by numbers of distinct cell types) [72, 73]. In apicomplexans, the splicing machinery appears to be largely conserved but features of gene structure such as intron number, length and distribution can be highly variable [18]. In our study, the difference in intron number between *T. gondii* and *P. falciparum* is a relevant example. Despite the differences, we found that alternative splicing for any given junction occurs at similar rates between the two parasites. This further supports the notion that the parasites share similar splicing processes. Intron number of genes is predicted to be positively associated with alternative splicing events [50, 74]. This is not simply a stochastic effect but rather related to the general decrease in 5' splice site strength with increasing intron number in many organisms [74]. As described above however, there is only a weak or moderately positive correlation between intron number and intron retention level of genes in the two parasites studied. Schmitz et al. [70] previously reported that other features such AT content and splice site entropy are important modulators of intron retention. This may also be the case in our study, given the observation that certain splice junctions are predisposed to retain their intron over others.

In addition to the difference in alternative splicing levels, there are differences in the composition of alternative splicing types between different organisms as reported by Kim and colleagues [71] and McGuire and colleagues [75]. For example, exon skipping is the predominant form of alternative splicing in metazoans, and intron retention the rarest [71]. Our analysis reveals that the opposite occurs in *T. gondii* and *P. falciparum*, with intron retention being the predominant event to occur, and exon skipping the rarest. This composition of alternative splicing type is similar to that observed for plants and fungi [71, 75, 76], though the reason is unclear. More recent studies find that intron retention has been previously under-detected in metazoans due to methodology limitations or confounding variables, but the high levels of exon skipping has been mostly undisputed [77]. Kim and colleagues [71] speculated that intron retention emerged as the earliest form of alternative splicing, before other mechanisms of complex splicing events evolved. There is some evidence for this, including the apparent shift towards increased exon skipping frequencies in early branching animals [77]. This is associated with the preservation of coding frames, suggesting a role of exon skipping in expanding proteome diversity [77]. In contrast to this, we found that the majority of the splicing events in *P. falciparum* and *T. gondii*, particularly intron retention, results in non-productive transcripts. Our results thus indicate that alternative splicing rarely contributes to generating diversity of protein sequence in these parasites, and may relate instead to transcriptomic complexity that impacts protein abundance. If that is true, a previous analysis that showed alternative splicing to be essential for *Plasmodium* stage differentiation [11] may possibly be explained by a requirement for modulation of abundance for specific proteins, rather than generation of protein sequences.

Consistent with our observation of alternative splicing playing a minor role in generating true proteome diversity in apicomplexans, many splicing events in other eukaryotes contribute little to the protein isoform repertoire. In particular, many transcripts contain premature termination codons (PTCs), at least in humans and yeast [78, 79]. Often, PTC transcripts are the result of the retention of intronic sequences that contain PTCs [80], but translational frameshifts from active splicing events such as alternative splice site selections have been similarly implicated [79, 81]. PTC transcripts are not normally translated but rather targeted for degradation through the nonsense-mediated decay (NMD) pathway [82, 83]. This is vital because the transcripts encode altered or truncated proteins which may exhibit deleterious activity [84]. Some studies

postulate that the predicted alternative splice events are the result of either experimental or transcriptional noise [85], or that a substantial portion of such transcripts are contaminating pre-mRNA molecules, and so do not represent true alternative splicing [86, 87]. Nevertheless, many RNA-seq-based analyses operate on the assumption that PTC transcripts are biologically significant or relevant [88]. Congruously, studies focusing on mature mRNA isoforms in other organisms suggest that non-productive transcripts mediate an additional layer of post-transcriptional regulation, through downstream RNA processing changes such as mRNA turnover, export, and microRNA silencing [54, 89, 90]. Alternative splicing in apicomplexans may also play a role in these processes. Strikingly, unique PTC transcript signatures are associated with distinct cell lineages [42, 91, 92] in multicellular eukaryotes, which may be analogous to the essential role of alternative splicing observed in stage-differentiation observed in *Plasmodium* [11].

Non-productive transcripts are typically degraded though the NMD pathway, and this has been shown to regulate gene expression at the post-transcriptional level [93]. However, it is difficult to conclusively define the function of the non-productive transcripts without experimentally testing the proteomic fates of these transcripts. In metazoans, non-productive transcripts often highlight genes that were downregulated following a transition of cellular states [92]. Our study is consistent with this, given the observation that the number of non-productive transcripts generally decreased with increasing transcript number. This, in association with NMD, has been shown to be crucial to the maintenance and differentiation of many cell types [94, 95]. In contrast, in organisms such *Paramecium tetraurelia,* non-productive transcripts appear to be the result of splicing error rather than function [96]. Regardless, because gene expression as measured by transcript levels do not necessarily translate to protein expression levels, our findings have potential implications for the interpretation of RNA-seq studies in these parasites. Several studies have already demonstrated the poor correlation between protein and mRNA expression profiles in apicomplexans [7, 8, 97]. Our results highlight that for many genes, raw quantifications of transcript abundance will correlate poorly with the number of copies of productive isoforms, and provides one source of mismatch between transcriptional initiation and protein abundance.

Genome annotation is a crucial element of RNA-seq data analysis. For *T. gondii* and *P. falciparum*, the task is a widely accomplished manual effort from experts in the research community. Although genome annotation was not the main focus of the study, the ONT datasets are able to reveal the structure of full-length transcripts. This is crucial in validating gene models. Our data are viewable through the ToxoDB and PlasmoDB web resources [98], and raw data are available at the Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA606986) which may aid the research community to further curate and validate the current annotations.

Conclusion

In this study, we have performed the first direct transcriptomic analyses on *T. gondii* and *P. falciparum.* We show that ONT direct RNA sequencing enables the quantification of gene expression despite a reduced throughput. In combination with the increased requirement for starting material, this means that the cost and time per bp sequenced remains higher than that of second-generation sequencing platforms. Nevertheless, because ONT direct RNA sequencing enables the detection of full-length transcripts without amplification, the tool remains promising for resolving the limitations of second-generation sequencing.

We demonstrated that alternative splicing is widespread in the two parasites, particularly intron retention. ONT direct RNA sequencing enabled us to determine the productivity of these

transcripts without complex computational methodologies, and we show that most the transcripts are premature terminating. This has implications for the quantification of gene expression, as it is highly unlikely for the wealth of transcript diversity that we identified to directly translate to protein isoforms.

Methods

Cell culture and RNA extraction

Toxoplasma gondii tachyzoites (Pru *Δku80*) were cultured on human foreskin fibroblasts in Dulbecco's Modified Eagle medium (DME) supplemented with 1% v/v fetal calf serum and 1% v/v Glutamax. Freshly egressed tachyzoites were washed, filter purified (5um) and collected for RNA extraction. *P. falciparum* (3D7) were cultured in complete media consisting of human erythrocytes (O+, 2% haematocrit), RPMI-HEPES, 5% (w/v) Albumax and 3.6% (w/v) sodium bicarbonate. We collected mixed stage parasite, purified from host RBCs via lysis with 0.05% (w/v) saponin, for RNA extraction. To obtain the 500ng of mRNA recommended for the library preparation, we used TRI Reagent (Sigma) for extraction of total RNA followed by the Dynabeads mRNA Purification Kit for polyadenylated (poly-A) mRNA (Thermo Scientific) purification according to the manufacturer's protocol. Purity and quantification of mRNA were determined via NanoDrop (Thermo Scientific) and a Qubit RNA HS Assay kit (Thermo Scientific).

Library prep and Nanopore sequencing

Libraries for the direct RNA sequencing were generated using the recommended protocol for the SQK-RNA001 kit (Oxford Nanopore Technologies). We loaded and sequenced the libraries on MinION R9.4 flow cells (Oxford Nanopore Technologies) for 48 hours. Base calling was performed concurrent with sequencing using Albacore (v 2.0), which was integrated within the MinION software (MinKNOW, v1.10.23). Only "pass" reads were used for subsequent analyses.

Mapping and qualitative analysis

ONT sequencing data was first checked for quality with FastQC (v.0.11.7) [99]. We then utilised Minimap2 (v. 2.1) [38] to map raw reads to the parasite genome and transcriptome from ToxoDB and PlasmoDB (r. 39), using the recommended preset commands. Intron length thresholds were set at 5000 and 1500 bases for *T. gondii* and *P. falciparum* respectively. Previously published Illumina datasets (SRR350746, ERR174301, ERR185969, ERR185970, ERR185971) were mapped using HISAT2 [100] using the preset commands. We checked for mapping quality with Samtools (v.1.7) [101], Picard (v.2.18.2) [102] and AlignQC (v.1.2) [103] . Further qualitative or quantitative analyses and graphical elements were done using the command-line interface and RStudio. We verified and illustrated subsets of mapped reads via IGV [104].

We correlated the ONT sequencing data with the Illumina datasets as previously described [105] using the wub package (v.0.2) [106]. The genome coverage of sequencing datasets were generated using bedtools genome coverage (v2.27) [107] and visualised via J-Circos (v1) [108]. Log2-fold ratios were calculated using DeepTools bamCompare (v.2.5.1) [109].

Alternative splicing analysis

We applied two approaches to analysing alternative splicing. We first identified intron retained junctions and transcripts using featureCounts (v.1.6.2) [48] on the genome mapped reads. featureCounts matches features specified in an annotation file (gff) to mapped reads. The annotation files used in the analyses were obtained from ToxoDB and PlasmoDB (r. 39), and

preprocessed via ToolShed (v.1.0) [110] to specifically extract intron coordinates and gene IDs. We set a minimum threshold requiring mapping to at least six-bases of the intron feature, and a minimum threshold of three reads mapping to the junction/transcript to be considered for further analysis. Percent Intron Retention (PIR) scores were calculated as the proportion of alternative splicing events to the sum of reads for each junction/gene. Productivity of full length transcripts was analysed using the Flair pipeline [51] using default parameters.

For the second approach, we used the junction_annotation.py script of RSeQC (v.2.6.4) [111] to identify novel or partial-novel junctions from the genome mapped reads based on the unmodified annotation file. Again, we filtered out junctions that had fewer than three supporting reads. The junctions were summarised into a table based on coordinate matching to the 5' and/or 3' of the expected canonical junction. We identified alternative 5'/3' splicing and exon skipping based on the coordinates and strandedness of junctions identified by RSeQC that were either consistent with or conflicted with the annotated junctions. We manually validated the data, matched junctions to available gene IDs, and again calculated the proportion of alternative splicing events to the sum of reads for each junction. Using the final dataset, we re-curated the list of intron-retained junctions to exclude for alternate 5'/3' splice changes. Proportional Venn diagrams were drawn using BioVenn [112].

Gene set enrichment analyses were carried out by ranking the genes based on their alternative splicing levels and using the first and third quartile of the ranked list as input for GO enrichment analysis via ToxoDB/PlasmoDB based on curated and computed assigned associations. We required the adjusted p-value to be smaller than 0.05 and FDR q-value of less than 0.25. This approach was validated using GSEA via WebGestalt [113].

Supplementary files

Supplementary table TS1. An Excel spreadsheet quantifying the expression of genes from Illumina and ONT data.

Supplementary table TS2. An Excel spreadsheet quantifying the levels of intron retention at the gene level.

Supplementary table TS3. An Excel spreadsheet summarising the GO enrichment results for genes with high and low levels of intron retention.

Supplementary table TS4_1. An Excel spreadsheet quantifying categories of alternative splicing at the junction level for *T. gondii.*

Supplementary table TS4_2. An Excel spreadsheet quantifying categories of alternative splicing at the junction level for *P. falciparum.*

Supplementary table TS5. An Excel spreadsheet listing previously and currently identified genes with alternative splicing excluding intron retention for *T. gondii.*

References

1. Cowman, A.F., et al., *Malaria: Biology and Disease.* Cell, 2016. **167**(3): p. 610-624.
2. Torgerson, P.R. and P. Mastroiacovo, *The global burden of congenital toxoplasmosis: a systematic review.* Bull World Health Organ, 2013. **91**(7): p. 501-8.
3. Furtado, J.M., et al., *Toxoplasmosis: A Global Threat.* Journal of Global Infectious Diseases, 2011. **3**(3): p. 281-284.
4. Horrocks, P., et al., *Control of gene expression in Plasmodium falciparum - ten years on.* Mol Biochem Parasitol, 2009. **164**(1): p. 9-25.
5. Wesseling, J.G., et al., *Stage-specific expression and genomic organization of the actin genes of the malaria parasite Plasmodium falciparum.* Mol Biochem Parasitol, 1989. **35**(2): p. 167-76.
6. Lopez-Barragan, M.J., et al., *Directional gene expression and antisense transcripts in sexual and asexual stages of Plasmodium falciparum.* BMC Genomics, 2011. **12**: p. 587.
7. Le Roch, K.G., et al., *Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle.* Genome Res, 2004. **14**(11): p. 2308-18.
8. Foth, B.J., et al., *Quantitative protein expression profiling reveals extensive post-transcriptional regulation and post-translational modifications in schizont-stage malaria parasites.* Genome Biol, 2008. **9**(12): p. R177.
9. Toenhake, C.G. and R. Bártfai, *What functional genomics has taught us about transcriptional regulation in malaria parasites.* Briefings in Functional Genomics, 2019.
10. Llora-Batlle, O., E. Tinto-Font, and A. Cortes, *Transcriptional variation in malaria parasites: why and how.* Brief Funct Genomics, 2019.
11. Yeoh, L.M., et al., *Alternative splicing is required for stage differentiation in malaria parasites.* Genome Biol, 2019. **20**(1): p. 151.
12. Mockenhaupt, S. and E.V. Makeyev, *Non-coding functions of alternative pre-mRNA splicing in development.* Seminars in cell & developmental biology, 2015. **47-48**: p. 32-39.
13. Neverov, A.D., et al., *Alternative splicing and protein function.* BMC Bioinformatics, 2005. **6**(1): p. 266.
14. Kelemen, O., et al., *Function of alternative splicing.* Gene, 2013. **514**(1): p. 1-30.
15. Birzele, F., G. Csaba, and R. Zimmer, *Alternative splicing and protein structure evolution.* Nucleic acids research, 2008. **36**(2): p. 550-558.
16. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.
17. Licatalosi, D.D. and R.B. Darnell, *RNA processing and its regulation: global insights into biological networks.* Nat Rev Genet, 2010. **11**(1): p. 75-87.
18. Yeoh, L.M., et al., *Alternative Splicing in Apicomplexan Parasites.* mBio, 2019. **10**(1): p. e02866-18.
19. van Dooren, G.G., et al., *Processing of an apicoplast leader sequence in Plasmodium falciparum, and the identification of a putative leader cleavage enzyme.* J Biol Chem, 2002. **277**(26): p. 23612-23619.
20. Pham, J.S., et al., *A dual-targeted aminoacyl-tRNA synthetase in Plasmodium falciparum charges cytosolic and apicoplast tRNACys.* Biochem J, 2014. **458**(3): p. 513-23.
21. Gadalla, N.B., et al., *Alternatively spliced transcripts and novel pseudogenes of the Plasmodium falciparum resistance-associated locus pfcrt detected in East African*

malaria patients. The Journal of antimicrobial chemotherapy, 2015. **70**(1): p. 116-123.
22. Jex, A.R., et al., *Genome and transcriptome of the porcine whipworm Trichuris suis.* Nature Genetics, 2014. **46**(7): p. 701-706.
23. Lees, J.G., J.A. Ranea, and C.A. Orengo, *Identifying and characterising key alternative splicing events in Drosophila development.* BMC genomics, 2015. **16**(1): p. 608-608.
24. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nat Rev Genet, 2016. **17**(6): p. 333-51.
25. Salzberg, S.L., et al., *GAGE: A critical evaluation of genome assemblies and assembly algorithms.* Genome research, 2012. **22**(3): p. 557-567.
26. Kozarewa, I., et al., *Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes.* Nature methods, 2009. **6**(4): p. 291-295.
27. Chaudhary, K., et al., *Differential localization of alternatively spliced hypoxanthine-xanthine-guanine phosphoribosyltransferase isoforms in Toxoplasma gondii.* J Biol Chem, 2005. **280**(23): p. 22053-9.
28. Jenjaroenpun, P., et al., *Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of Saccharomyces cerevisiae CEN.PK113-7D.* Nucleic Acids Research, 2018. **46**(7): p. e38-e38.
29. Jain, M., et al., *Nanopore sequencing and assembly of a human genome with ultra-long reads.* Nature Biotechnology, 2018. **36**: p. 338.
30. Byrne, A., et al., *Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells.* Nature Communications, 2017. **8**: p. 16027.
31. Simpson, J.T., et al., *Detecting DNA cytosine methylation using nanopore sequencing.* Nature Methods, 2017. **14**: p. 407.
32. Workman, R.E., et al., *Nanopore native RNA sequencing of a human poly(A) transcriptome.* bioRxiv, 2018: p. 459529.
33. Chappell, L., et al., *Refining the transcriptome of the human malaria parasite Plasmodium falciparum using amplification-free RNA-seq.* bioRxiv, 2019: p. 852038.
34. Garalde, D.R., et al., *Highly parallel direct RNA sequencing on an array of nanopores.* Nature Methods, 2018. **15**(3): p. 201-206.
35. Viehweger, A., *Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis.* Genome Research, 2019. **29**(9): p. 1545.
36. Sharon, D., et al., *A single-molecule long-read survey of the human transcriptome.* Nature biotechnology, 2013. **31**(11): p. 1009-1014.
37. Soneson, C., et al., *A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes.* Nature Communications, 2019. **10**(1): p. 3359.
38. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094-3100.
39. Shaw, P.J., et al., *Estimating mRNA lengths from Plasmodium falciparum genes by Virtual Northern RNA-seq analysis.* Int J Parasitol, 2016. **46**(1): p. 7-12.
40. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing.* Nature, 2010. **463**(7280): p. 457-63.
41. Braunschweig, U., et al., *Widespread intron retention in mammals functionally tunes transcriptomes.* Genome Res, 2014. **24**(11): p. 1774-86.

42. Wong, J.J., et al., *Orchestrated intron retention regulates normal granulocyte differentiation.* Cell, 2013. **154**(3): p. 583-95.
43. Boutz, P.L., A. Bhutkar, and P.A. Sharp, *Detained introns are a novel, widespread class of post-transcriptionally spliced introns.* Genes & development, 2015. **29**(1): p. 63-80.
44. Sakharkar, M.K., V.T. Chow, and P. Kangueane, *Distributions of exons and introns in the human genome.* In Silico Biol, 2004. **4**(4): p. 387-93.
45. Ivashchenko, A.T., V.A. Khailenko, and A. Atambaeva Sh, *[Variations of the length of exons and introns in human genome genes].* Genetika, 2009. **45**(1): p. 22-9.
46. Lau, Y.L., et al., *Deciphering the Draft Genome of Toxoplasma gondii RH Strain.* PLoS One, 2016. **11**(6): p. e0157901.
47. Hall, N., et al., *Sequence of Plasmodium falciparum chromosomes 1, 3-9 and 13.* Nature, 2002. **419**(6906): p. 527-31.
48. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2014. **30**(7): p. 923-30.
49. Gonzalez-Porta, M., et al., *Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene.* Genome Biol, 2013. **14**(7): p. R70.
50. Schmitz, U., et al., *Intron retention enhances gene regulatory complexity in vertebrates.* Genome biology, 2017. **18**(1): p. 216-216.
51. Tang, A.D., et al., *Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns.* bioRxiv, 2018: p. 410183.
52. Nagy, E. and L.E. Maquat, *A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance.* Trends Biochem Sci, 1998. **23**(6): p. 198-9.
53. Ge, Y. and B.T. Porse, *The functional consequences of intron retention: alternative splicing coupled to NMD as a regulator of gene expression.* Bioessays, 2014. **36**(3): p. 236-43.
54. Braunschweig, U., et al., *Widespread intron retention in mammals functionally tunes transcriptomes.* Genome research, 2014. **24**(11): p. 1774-1786.
55. Neverov, A.D., et al., *Alternative splicing and protein function.* BMC bioinformatics, 2005. **6**: p. 266-266.
56. Sidik, S.M., et al., *A Genome-wide CRISPR Screen in Toxoplasma Identifies Essential Apicomplexan Genes.* Cell, 2016. **166**(6): p. 1423-1435.e12.
57. Zhang, M., et al., *Uncovering the essential genes of the human malaria parasite Plasmodium falciparum by saturation mutagenesis.* Science, 2018. **360**(6388).
58. Pai, A.A., et al., *The kinetics of pre-mRNA splicing in the Drosophila genome and the influence of gene architecture.* Elife, 2017. **6**.
59. Skinner, S.O., et al., *Single-cell analysis of transcription kinetics across the cell cycle.* Elife, 2016. **5**: p. e12175.
60. Yeoh, L.M., et al., *A serine–arginine-rich (SR) splicing factor modulates alternative splicing of over a thousand genes in Toxoplasma gondii.* Nucleic Acids Research, 2015. **43**(9): p. 4661-4675.
61. Tsai, K.-W., et al., *Sequence features involved in the mechanism of 3' splice junction wobbling.* BMC molecular biology, 2010. **11**: p. 34-34.
62. Jenjaroenpun, P., et al., *Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of Saccharomyces cerevisiae CEN.PK113-7D.* Nucleic Acids Res, 2018. **46**(7): p. e38.

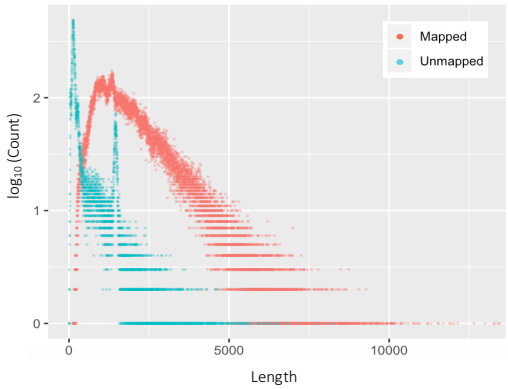
63. Moldován, N., et al., *Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus.* Scientific Reports, 2018. **8**(1): p. 8604.
64. Dozmorov, M.G., et al., *Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data.* BMC bioinformatics, 2015. **16 Suppl 13**(Suppl 13): p. S10-S10.
65. Gardner, M.J., et al., *Genome sequence of the human malaria parasite Plasmodium falciparum.* Nature, 2002. **419**(6906): p. 498-511.
66. Runtuwene, L.R., et al., *Nanopore sequencing of drug-resistance-associated genes in malaria parasites, Plasmodium falciparum.* Scientific Reports, 2018. **8**(1): p. 8286.
67. Loven, J., et al., *Revisiting global gene expression analysis.* Cell, 2012. **151**(3): p. 476-82.
68. Tarazona, S., et al., *Differential expression in RNA-seq: a matter of depth.* Genome Res, 2011. **21**(12): p. 2213-23.
69. Barbosa-Morais, N.L., et al., *The Evolutionary Landscape of Alternative Splicing in Vertebrate Species.* Science, 2012. **338**(6114): p. 1587.
70. Schmitz, U., et al., *Intron retention enhances gene regulatory complexity in vertebrates.* Genome Biology, 2017. **18**(1): p. 216.
71. Kim, E., A. Magen, and G. Ast, *Different levels of alternative splicing among eukaryotes.* Nucleic acids research, 2007. **35**(1): p. 125-131.
72. Lynch, M. and J.S. Conery, *The Origins of Genome Complexity.* Science, 2003. **302**(5649): p. 1401.
73. Chen, L., et al., *Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity.* Molecular biology and evolution, 2014. **31**(6): p. 1402-1413.
74. Irimia, M., D. Penny, and S.W. Roy, *Coevolution of genomic intron number and splice sites.* Trends Genet, 2007. **23**(7): p. 321-5.
75. McGuire, A.M., et al., *Cross-kingdom patterns of alternative splicing and splice recognition.* Genome Biology, 2008. **9**(3): p. R50.
76. Sieber, P., et al., *Comparative Study on Alternative Splicing in Human Fungal Pathogens Suggests Its Involvement During Host Invasion.* Frontiers in Microbiology, 2018. **9**(2313).
77. Grau-Bové, X., I. Ruiz-Trillo, and M. Irimia, *Origin of exon skipping-rich transcriptomes in animals driven by evolution of gene architecture.* Genome biology, 2018. **19**(1): p. 135-135.
78. Kawashima, T., et al., *Widespread use of non-productive alternative splice sites in Saccharomyces cerevisiae.* PLoS Genet, 2014. **10**(4): p. e1004249.
79. Lewis, B.P., R.E. Green, and S.E. Brenner, *Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans.* Proc Natl Acad Sci U S A, 2003. **100**(1): p. 189-92.
80. Sayani, S., et al., *Widespread impact of nonsense-mediated mRNA decay on the yeast intronome.* Mol Cell, 2008. **31**(3): p. 360-70.
81. Plass, M., et al., *RNA secondary structure mediates alternative 3'ss selection in Saccharomyces cerevisiae.* RNA (New York, N.Y.), 2012. **18**(6): p. 1103-1115.
82. Weischenfeldt, J., et al., *NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements.* Genes & development, 2008. **22**(10): p. 1381-1396.
83. Hwang, J. and L.E. Maquat, *Nonsense-mediated mRNA decay (NMD) in animal embryogenesis: to die or not to die, that is the question.* Current opinion in genetics & development, 2011. **21**(4): p. 422-430.

84. Shi, M., et al., *Premature Termination Codons Are Recognized in the Nucleus in A Reading-Frame Dependent Manner.* Cell Discov, 2015. **1**.
85. Harati, S., J.H. Phan, and M.D. Wang, *Investigation of factors affecting RNA-seq gene expression calls.* Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference, 2014. **2014**: p. 5232-5235.
86. Zhao, S., et al., *Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion.* Scientific reports, 2018. **8**(1): p. 4781-4781.
87. Ge, Y. and B.T. Porse, *The functional consequences of intron retention: Alternative splicing coupled to NMD as a regulator of gene expression.* BioEssays, 2014. **36**(3): p. 236-243.
88. Zhang, A.Y., et al., *A data-driven approach to characterising intron signal in RNA-seq data.* bioRxiv, 2018: p. 352823.
89. Yap, K. and E.V. Makeyev, *Regulation of gene expression in mammalian nervous system through alternative pre-mRNA splicing coupled with RNA quality control mechanisms.* Mol Cell Neurosci, 2013. **56**: p. 420-8.
90. Wang, M., A.T. Branco, and B. Lemos, *The Y Chromosome Modulates Splicing and Sex-Biased Intron Retention Rates in Drosophila.* Genetics, 2018. **208**(3): p. 1057-1067.
91. Edwards, C.R., et al., *A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages.* Blood, 2016. **127**(17): p. e24.
92. Middleton, R., et al., *IRFinder: assessing the impact of intron retention on mammalian gene expression.* Genome Biology, 2017. **18**(1): p. 51.
93. Nickless, A., J.M. Bailis, and Z. You, *Control of gene expression through the nonsense-mediated RNA decay pathway.* Cell & bioscience, 2017. **7**: p. 26-26.
94. Han, X., et al., *Nonsense-mediated mRNA decay: a 'nonsense' pathway makes sense in stem cell biology.* Nucleic acids research, 2018. **46**(3): p. 1038-1051.
95. Vanichkina, D.P., et al., *Challenges in defining the role of intron retention in normal biology and disease.* Seminars in Cell & Developmental Biology, 2018. **75**: p. 40-49.
96. Saudemont, B., et al., *The fitness cost of mis-splicing is the main determinant of alternative splicing patterns.* Genome biology, 2017. **18**(1): p. 208-208.
97. Foth, B.J., et al., *Quantitative time-course profiling of parasite and host cell proteins in the human malaria parasite Plasmodium falciparum.* Mol Cell Proteomics, 2011. **10**(8): p. M110.006411.
98. Aurrecoechea, C., et al., *EuPathDB: the eukaryotic pathogen genomics database resource.* Nucleic Acids Res, 2017. **45**(D1): p. D581-d591.
99. Andrews, S., *FastQC A Quality Control tool for High Throughput Sequence Data*. 2010: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
100. Kim, D., et al., *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.* Nature Biotechnology, 2019. **37**(8): p. 907-915.
101. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
102. Broad Institute, *Picard Toolkit*. 2019: http://broadinstitute.github.io/picard/.
103. Weirather, J.L., et al., *Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis.* F1000Research, 2017. **6**: p. 100-100.
104. Robinson, J.T., et al., *Integrative genomics viewer.* Nature biotechnology, 2011. **29**(1): p. 24-26.

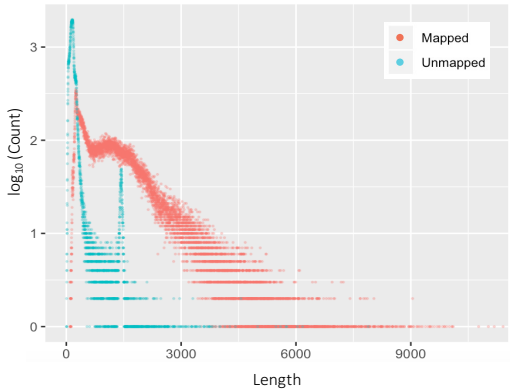
105. Garalde, D.R., et al., *Highly parallel direct RNA sequencing on an array of nanopores.* Nat Methods, 2018. **15**(3): p. 201-206.
106. Oxford Nanopore Technologies Ltd., *Wub*. 2016: https://github.com/nanoporetech/wub.
107. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-842.
108. An, J., et al., *J-Circos: an interactive Circos plotter.* Bioinformatics, 2015. **31**(9): p. 1463-5.
109. Ramírez, F., et al., *deepTools: a flexible platform for exploring deep-sequencing data.* Nucleic acids research, 2014. **42**(Web Server issue): p. W187-W191.
110. Blankenberg, D., et al., *Dissemination of scientific software with Galaxy ToolShed.* Genome Biology, 2014. **15**(2): p. 403.
111. Wang, L., S. Wang, and W. Li, *RSeQC: quality control of RNA-seq experiments.* Bioinformatics, 2012. **28**(16): p. 2184-5.
112. Hulsen, T., J. de Vlieg, and W. Alkema, *BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams.* BMC Genomics, 2008. **9**(1): p. 488.
113. Liao, Y., et al., *WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs.* Nucleic Acids Res, 2019. **47**(W1): p. W199-w205.

A

T. gondii

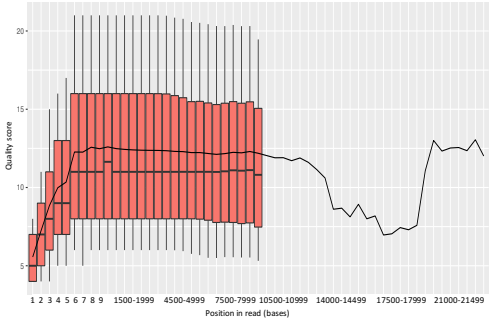


P. falciparum

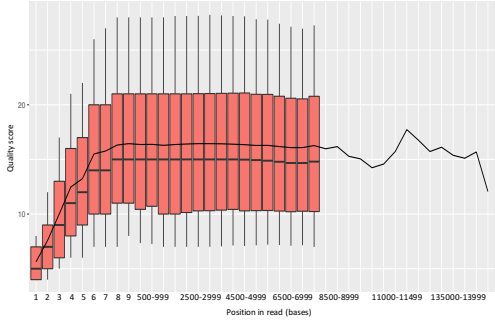


B

T. gondii

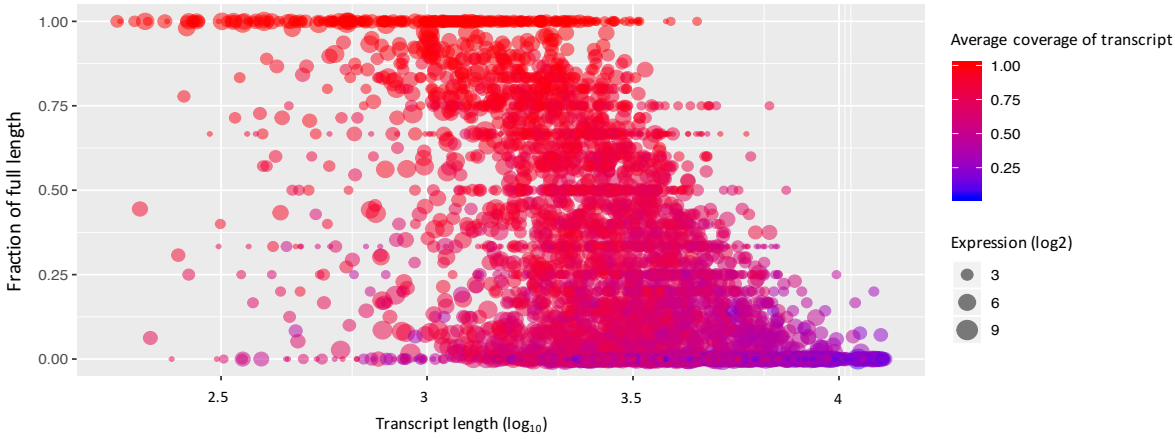


P. falciparum



C

T. gondii



P. falciparum

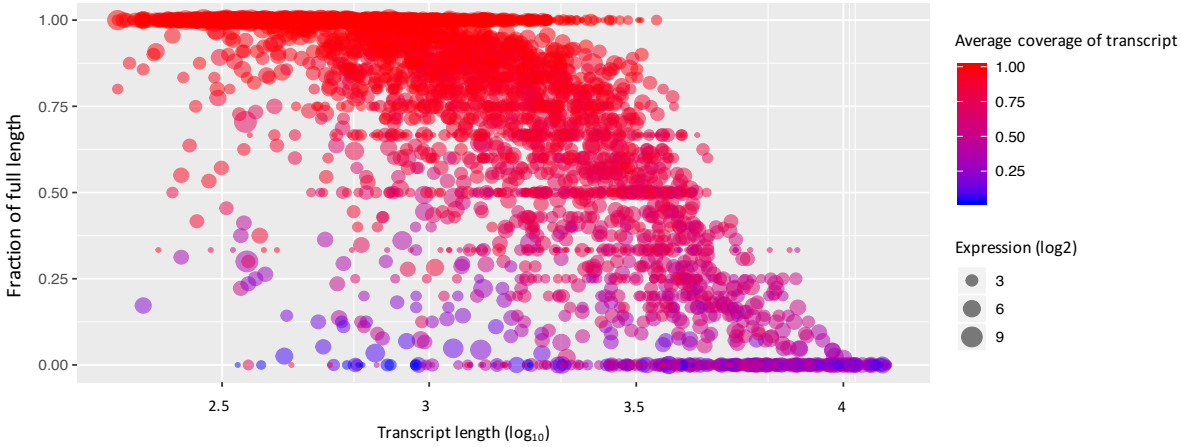
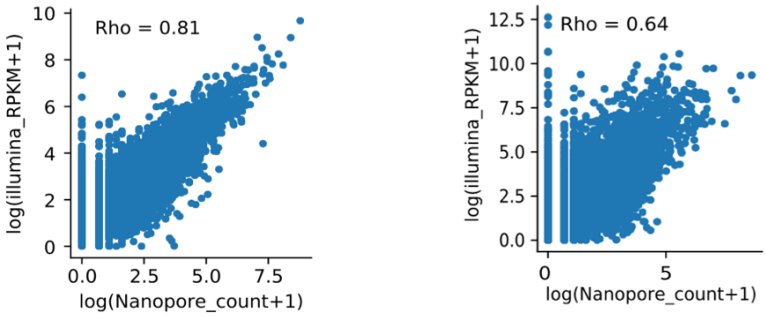


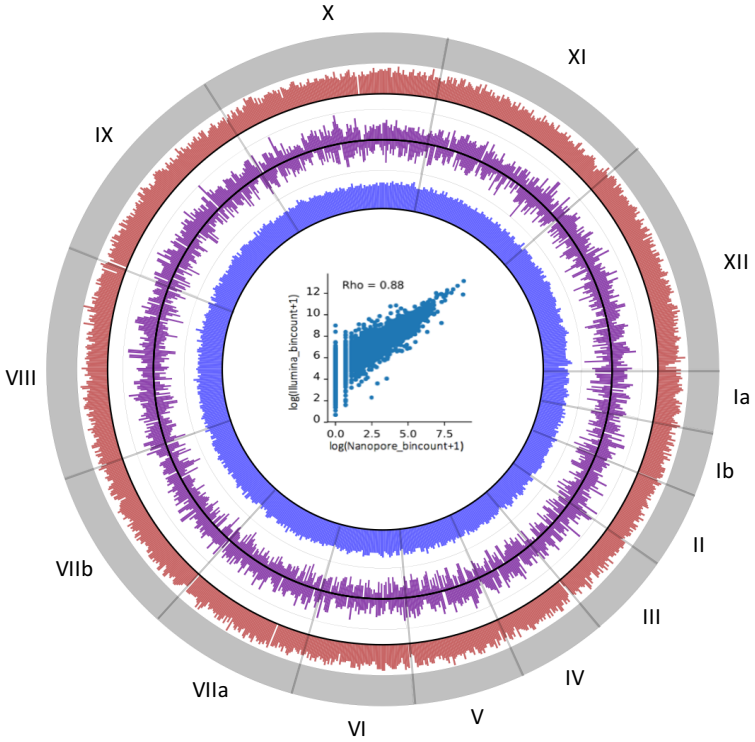
Figure 1. Summary of the ONT direct RNA sequencing data from *T. gondii* and *P. falciparum*. (A) Scatterplot of read length distribution of mapped (red) and unmapped (blue) reads. (B) Boxplot of quality scores across all bases at each position of the mapped sequencing reads. (C) Bubble scatter heat plots of the fraction of full length transcripts against transcript length. Size and color denotes expression and average coverage respectively.

A



B

T. gondii



P. falciparum

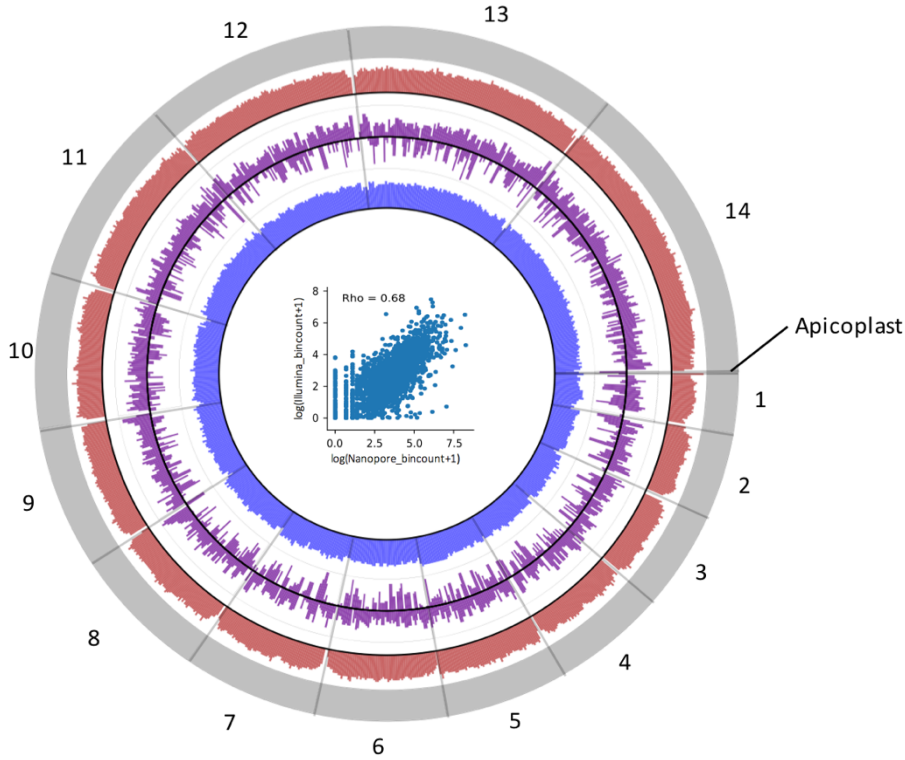


Figure 2. Comparisons between ONT direct RNA sequencing and Illumina datasets. (A) Correlation between trancriptome mapped read counts for *T. gondii* (left) and *P. falciparum* (right) as presented as a scatterplot. The Spearman correlation coefficient is shown. (B) Circos plots of genome mapped reads. Outer band (grey) represents the reference genome/chromosomes. The red and blue bands represent the genome coverage of ONT direct RNA and Illumina reads respectively. The purple band is the log2 fold change between the two datasets. The scatterplots within the circos plots show the correlation between the genome mapped read bin counts.

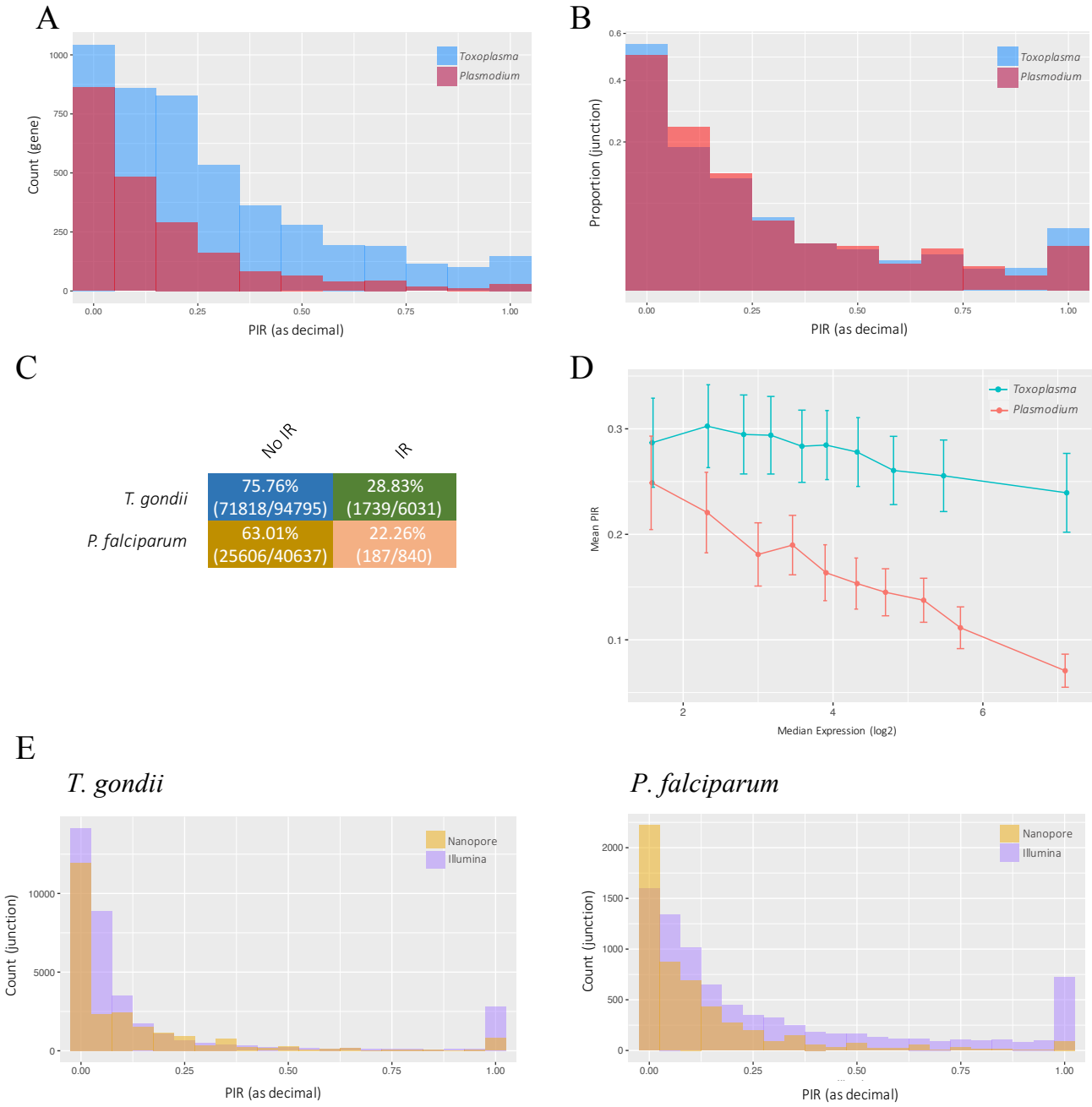
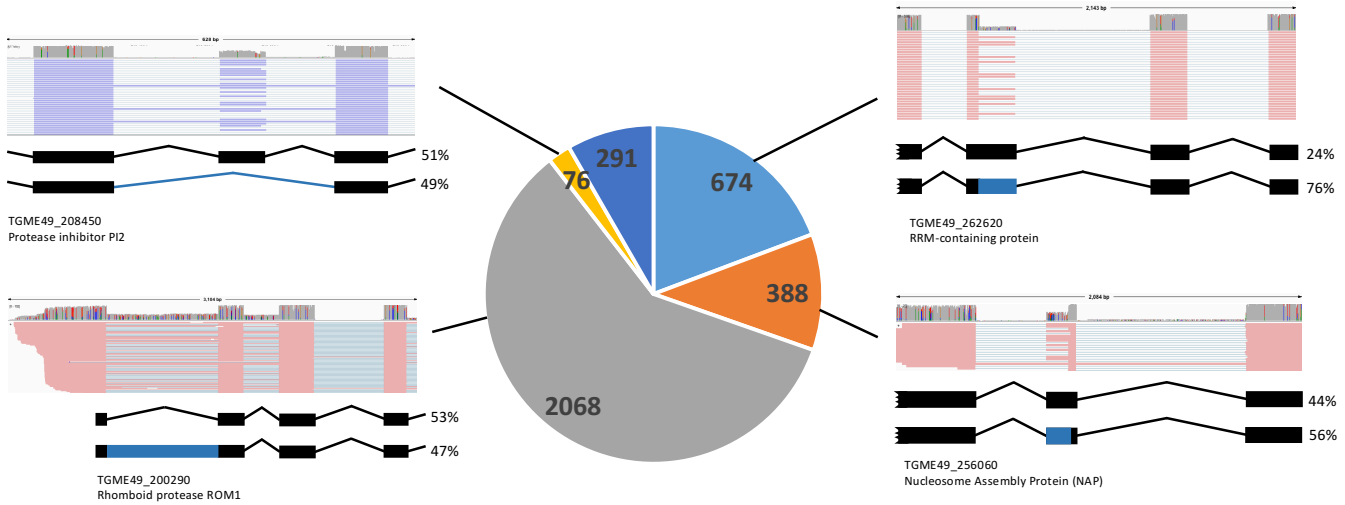
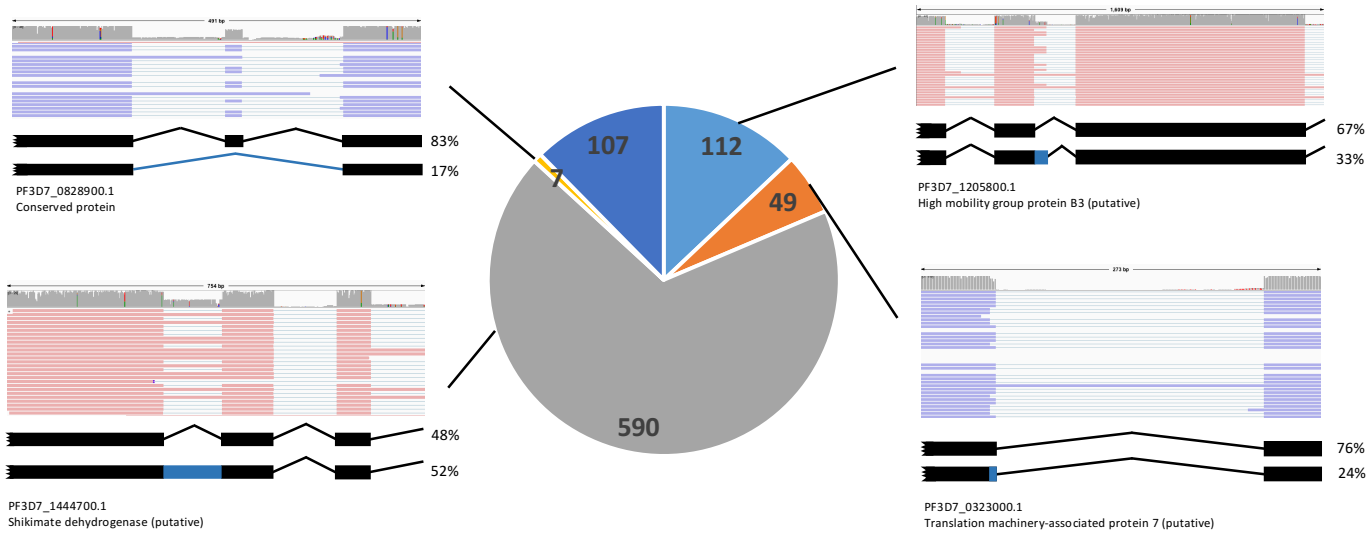


Figure 3. Analysis of intron retention using ONT direct RNA sequencing reads. Levels of intron retention are represented as Percent Intron Retention (PIR). (A) Distribution of intron retention levels over each gene as represented as total counts. (B) Distribution of intron retention levels over each junction as represented as the proportion of each bin count over the sum all intronic junctions. (C) Table of transcript productivity based on intron retention events as analysed using the FLAIR pipeline. Numbers in bracket are the transcript counts (D) Relationship between levels of intron retention and gene expression. Genes were classified into 10 bins of equal number based on expression. The median expression of the genes in each bin is used to represent expression for each bin. Error bars represent the 95% CI. (E) Comparisons of intron retention quantification between ONT and Illumina datasets.

T. gondii



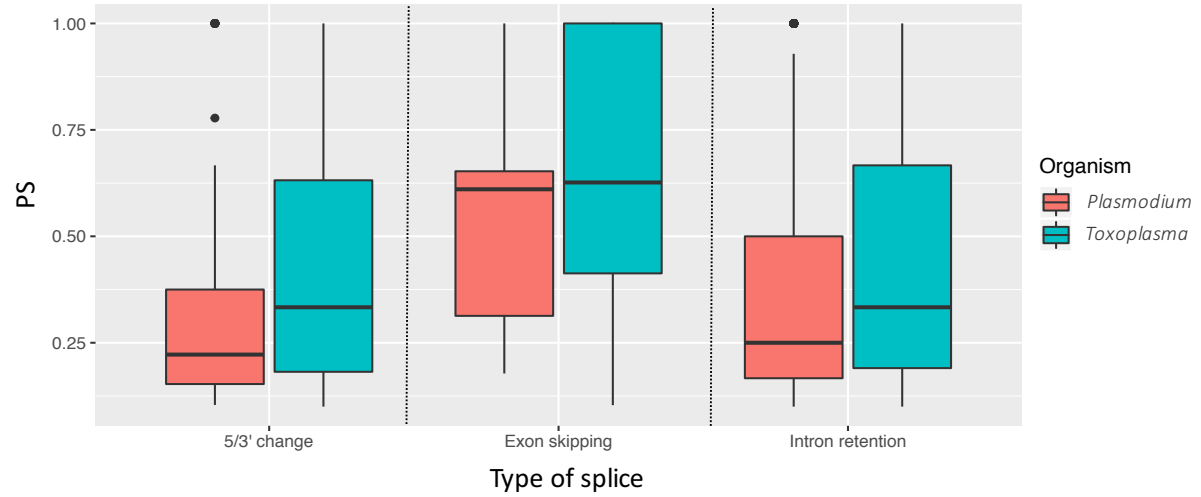
P. falciparum



5' change 3' change Intron retention Exon skipping Uncategorised

Figure 4. Summary analysis of alternative splicing from the ONT direct RNA sequencing datasets. Pie charts show the number, proportion and categorisation of genes with alternatively spliced transcripts equaling or exceeding 10% of its total transcript. An example of each event is presented. Red and blue represents sense and anti-sense transcripts respectively.

A



B

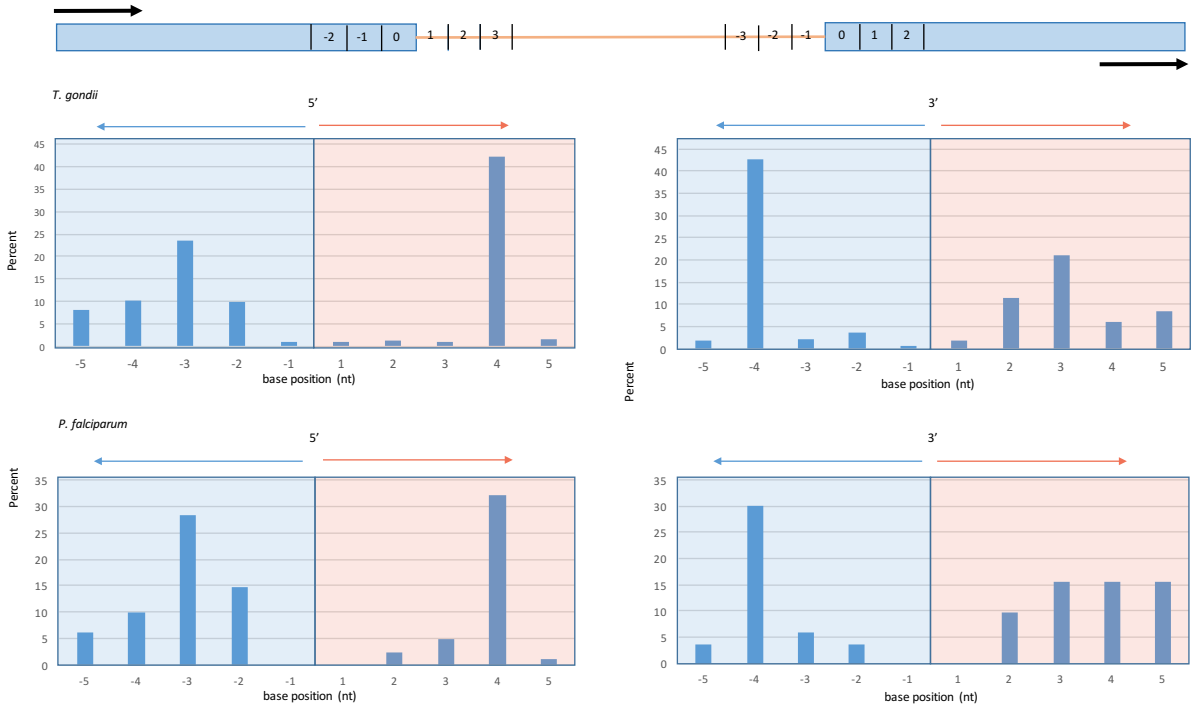
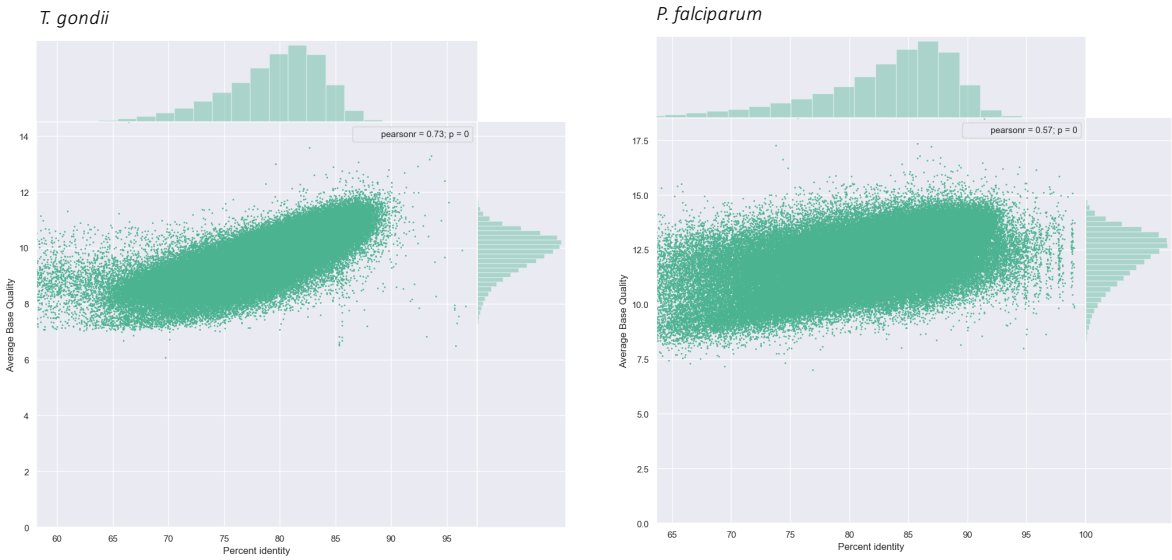
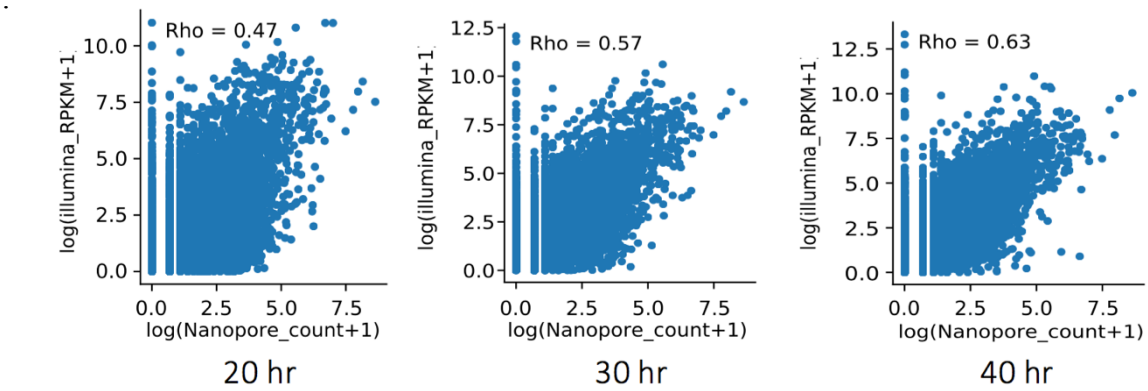


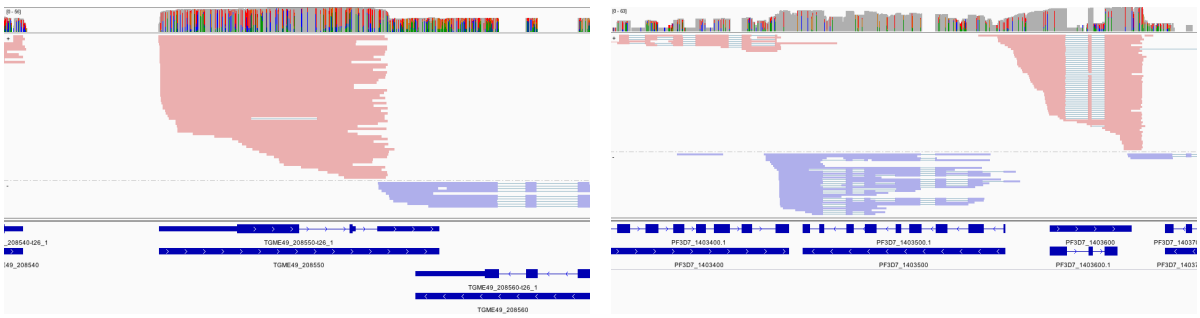
Figure 5. (A) Levels of each major alternative splice type as represented as Percent Splice (PS). (B) Distribution of alternate 5'/3' splice positions within 5 bases to the dominant splice site.



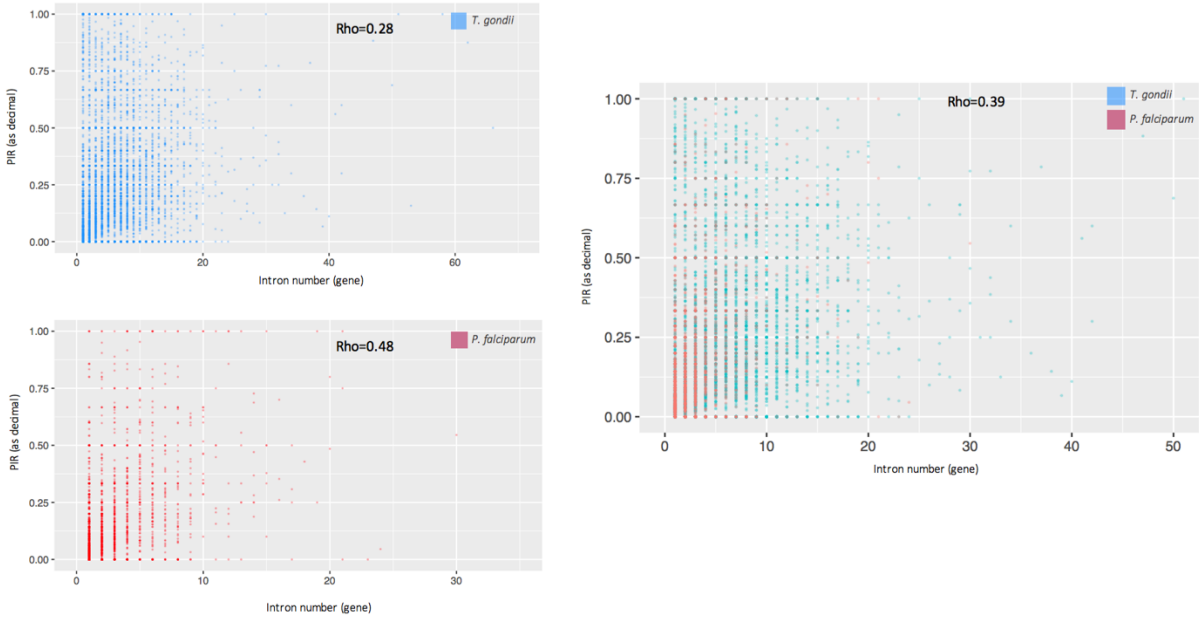
Supplementary Figure S1. Scatterplot shows the correlation between transcript identity (%) and average quality of the reads.



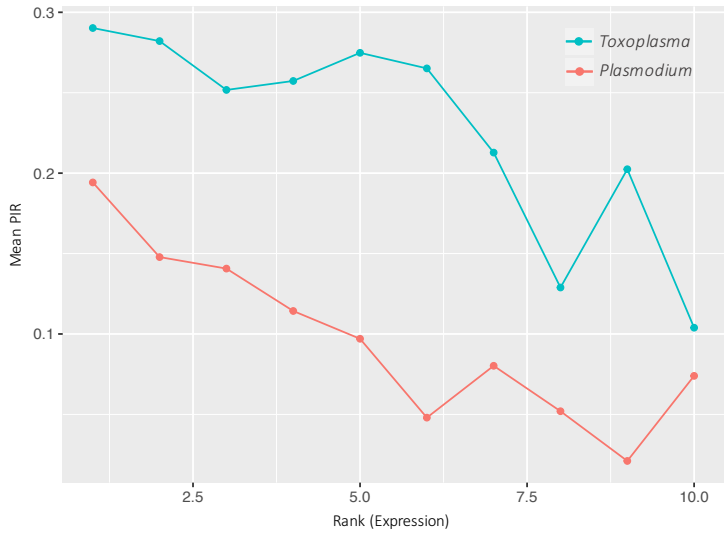
Supplementary Figure S2. Scatterplot shows the correlation between trancriptome mapped read counts for ONT direct RNA sequencing from mixed stage parasites and Illumina datasets from three developmental time points of *P. falciparum*. The Spearman correlation coefficient is shown.



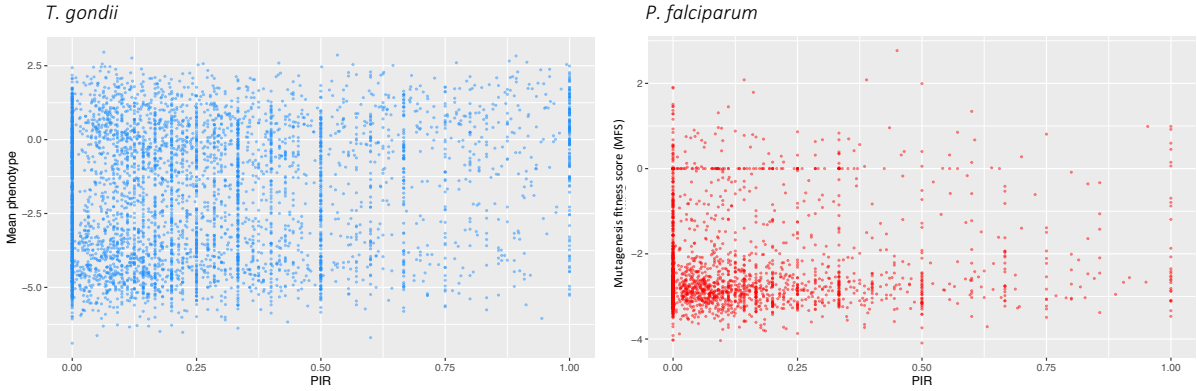
Supplementary Figure S3. IGV snapshots of genes which were flagged as having full intron retention and do not appear to conform with the gene model. Left: *T. gondii*; Right: *P. falciparum*.



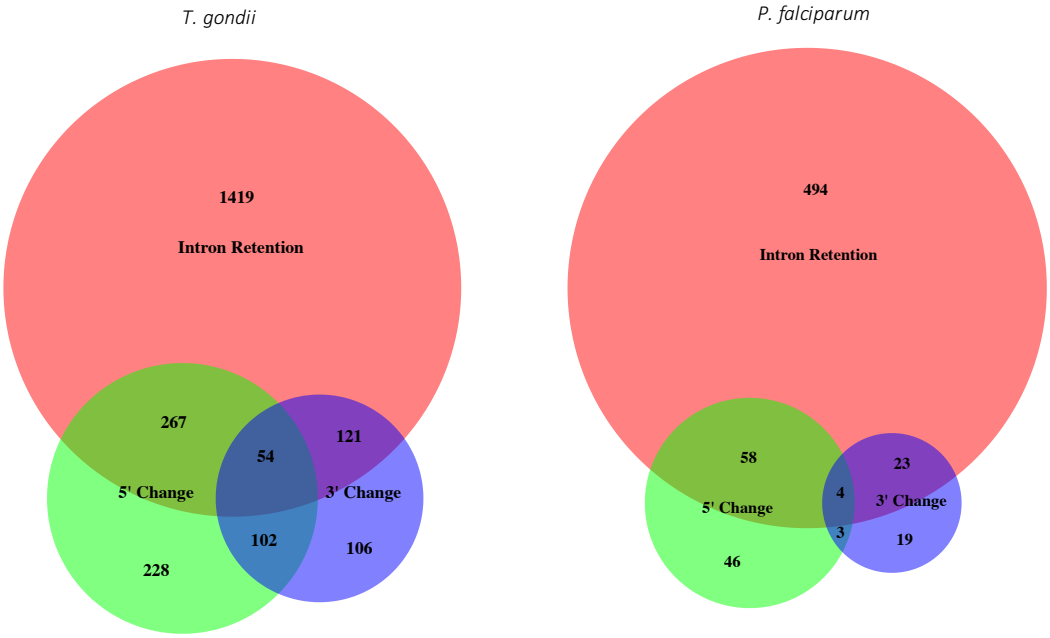
Supplementary Figure S4. Scatterplots show the correlation between intron retention levels and intron number per gene. The Spearman correlation coefficient is shown.



Supplementary Figure S5. Relationship between levels of intron retention and gene expression. Introns were classified into 10 bins of equal read number based on expression (1 = lowest, 10 = highest). Levels of intron retention were calculated as the proportion of intron retained reads over all reads from each bin.



Supplementary Figure S6. Scatterplots show the relationship between intron retention levels and gene essentiality. Essentiality is represented as mean phenotype and mutagenesis fitness score for *T. gondii* and *P. falciparum* respectively



Supplementary Figure S7. Venn diagrams display the number of genes with exclusive or overlapping intron retention, alternative 5' splicing and alternative 3' splicing events.